Image Retagging*

Dong Liu
School of Computer Sci.& Tec.
Harbin Institute of Technology
Harbin 150001, P. R. China
dongliu.hit@gmail.com

Xian-Sheng Hua
Microsoft Research Asia
49 Zhichun Road
Beijing 100080, P. R. China
xshua@microsoft.com

Meng Wang
Microsoft Research Asia
49 Zhichun Road
Beijing 100080, P. R. China
mengwang@microsoft.com

Hong-Jiang Zhang
Microsoft Adv. Tech. Center
49 Zhichun Road
Beijing 100080, P. R. China
hjzhang@microsoft.com

ABSTRACT

Online social media repositories such as Flickr and Zooomr allow users to manually annotate their images with freely-chosen tags, which are then used as indexing keywords to facilitate image search and other applications. However, these tags are frequently imprecise and incomplete, though they are provided by human beings, and many of them are almost only meaningful for the image owners (such as the name of a dog). Thus there is still a gap between these tags and the actual content of the images, and this significantly limits tag-based applications, such as search and browsing. To tackle this issue, this paper proposes a social image "retagging" scheme that aims at assigning images with better content descriptors. The refining process, including denoising and enriching, is formulated as an optimization framework based on the consistency between "visual similarity" and "semantic similarity" in social images, that is, the visually similar images tend to have similar semantic descriptors, and vice versa. An effective iterative bound optimization algorithm is applied to learn the improved tag assignment. In addition, as many tags are intrinsically not closely-related to the visual content of the images, we employ a knowledge based method to differentiate visual content related tags from unrelated ones and then constrain the tagging vocabulary of our automatic algorithm within the content related tags. Finally, to improve the coverage of the tags, we further enrich the tag set with appropriate synonyms and hypernyms based on an external knowledge base. Experimental results on a Flickr image collection demonstrate the effectiveness of this approach. We will also show the remarkable performance improvements brought by retagging via two applications, i.e., tag-based search and automatic annotation.

*This work was performed at Microsoft Research Asia.

MM'10, October 25–29, 2010, Firenze, Italy.

Categories and Subject Descriptors

H.3.1 [**Information Storage And Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Experimentation, Performance

Keywords

Image Tagging, Image Search, Tag Refinement, Retagging

1. INTRODUCTION

With the advent of Web 2.0 technology, there is an explosion of social media sharing system available online (e.g., Flickr, Youtube and Zooomr). Rather than simply searching for and passively viewing media content, such media repositories allow users to upload their media data and annotate them with freely-chosen tags, and this underscores a transformation of the Web as fundamental as its birth [1]. As one of the emerging Web 2.0 activities, tagging becomes a more and more frequently-applied means to organize, index and search media content for general users, and it provides a potential way to realize real large-scale content-based multimedia retrieval [2].

Despite the high popularity of tagging social images manually, the quality of tags is actually far from satisfactory as they are freely entered by grassroots Internet users. For example, many users make slips in spelling when entering tags and some users choose "batch tagging"[1] for efficiency which may introduce many noises. In addition, user-provided tags are often biased towards personal perspectives and context cues [3, 4], and thus there is a gap between these tags and the content of the images that common users are interested in. For example, an image uploader may tag his dog photos with "bomb", and it may make these photos appear in the search results of query "bomb". On the other hand, many potentially useful tags may be missed, as it is impractical for average users to annotate the images comprehensively. To summarize, user-provided tags are often *imprecise*, *biased* and *incomplete* for describing the content of the images.

As an example, Figure 1 illustrates an image from Flickr and its associated tags. From the figure we can observe that only "kitty" and "animal" truly describe the visual content

[1]Flickr provides a "batch tagging" feature that can simultaneously assign tags for a batch of images.

Figure 1: An example image from Flickr and its associated tags. The tags marked in red color truly describe the content of the image and the other tags are content-unrelated or irrelevant.

of the image, and the other tags are imprecise or subjective. Meanwhile, some other tags that should be used to describe the visual content are missed, such as "grass" and "flower". Moreover, if we further consider the concepts' lexical variability and the hierarchy of semantic expressions, the tags such as "kitten", "pussy" and "feline" also need to be added.

The *imprecise*, *biased* and *incomplete* characteristics of the tags have significantly limited the performance of social image search and organization. For example, they will degrade precision and recall rates in tag-based image search. This paper is aiming at improving the quality of the tags (in terms of describing the real content of images), thus the performance of tag-based applications can be greatly improved.

The most straightforward approach to tackle the difficulty is to ask humans to check the tags associated with each image, but obviously this way is infeasible considering the large number of images and tags. Therefore, an automatic method based on the available information, such as the content of images and their existing tags, is desired. In this work, we propose an optimization framework to improve the tag quality based on the following two observations in real-world social images. First, consistency between visual similarity and semantic similarity, that is, similar images often reflect similar semantic theme, and thus are annotated with similar tags. Second, user-provided tags, despite imperfect, still reveal the primary semantic theme of the image content. In our approach, the improved tag assignments are automatically discovered by maximizing the consistency between visual similarity and semantic similarity while minimizing the deviation from initially user-provided tags. This is actually mining information from different channels to complement each other in a collective way.

However, the first observation mentioned above is mainly applicable for "content related" tags[2]. That is, those tags that have high correspondence with the visual content. If we introduce "content unrelated" tags into the above optimization framework, we may even degrade the performance of the scheme. This is actually also one of the main difficulties for all automatic learning based image annotation approaches [5, 6], as well as one of the main reasons causing "semantic gaps" except for representative features and learning models. Accordingly, we propose a method to filter out those content unrelated tags to ensure that the quality of content related tags can be significantly improved.

Though the above tag refinement approach is able to add some new tags to an image (for example, a tag of an image may be transferred into a similar image originally without this tag), it is not able to "invent" new tags for a particular image if the tags are not assigned to its similar images. Therefore, as a post-processing step, we propose to use an external knowledge base to further enrich the tags to a certain extent, that is, adding synonyms and hypernyms[3].

1.1 Overview of the Proposed Scheme

As aforementioned, the goal of the image retagging scheme is to bridge the gap between tags and the content of images[4]. As illustrated in Figure 2, The scheme consists of three components: tag filtering, tag refinement, and further enrichment. Tag filtering eliminates content-unrelated tags based on a vocabulary that is built with both WordNet lexical knowledge base [7] and domain knowledge in vision field. With the obtained content-related tags, we then perform a refinement on them by modeling the consistency of visual and semantic similarities between images in the tag refinement component. Finally, the further enrichment component expands each tag with appropriate synonyms and hypernyms by mining the WordNet lexical knowledge base as well as the statistic information on social image website. The entire approach is fully automatic without needing any user interaction. It is also worth noting that the proposed retagging scheme can not only modify the tags but also discover their confidence (or relevance) scores with respect to their associated images, which will be useful in many tag-based applications [8, 9, 10].

1.2 Contribution of this work

The main contributions of this paper can be summarized as follows:

- We propose an optimization framework to improve the quality of tags by modeling the consistency of visual and semantic similarities of images. We also introduce an efficient iterative algorithm to find its solution.
- Propose a method to construct a content-related vocabulary based on both lexical and domain knowledge, which can be used to differentiate content-unrelated tags and content-related ones.
- We investigate how to mine lexical knowledge and statistical information on social image website to expand each tag with its synonyms and hypernyms.

2. RELATED WORK

Ames et al. [11] have explored the motivation of tagging on the Flickr website and they claim that most users tag images to make them better accessible to the general public.

[2]In this work, we regard whether or not a tag is content-related as its **intrinsic** property (in Section 3 we will discuss it in more detail). For example, the tag "cool" in Figure 1 is content-unrelated. In contrary, we call a tag irrelevant with respect to an image if it does not describe the image's visual content, such as the "dog" and "boy" in Figure 1. Thus, content-related tags can also be irrelevant.

[3]In search, this issue can also be solved by "query expansion", that is, expanding query term with its synonyms and hyponyms. But enriching image tags will be useful for other tag-based applications as well, such as image organizing and automatic annotating untagged images.

[4]One may argue that user-provided tags will always be useful (at least for the owners themselves). But making tags better describe the content of images will benefit much more users in tag-based applications. In addition, we can also choose to keep both the original tags and the ones after retagging to support content owners and common users at the same time.

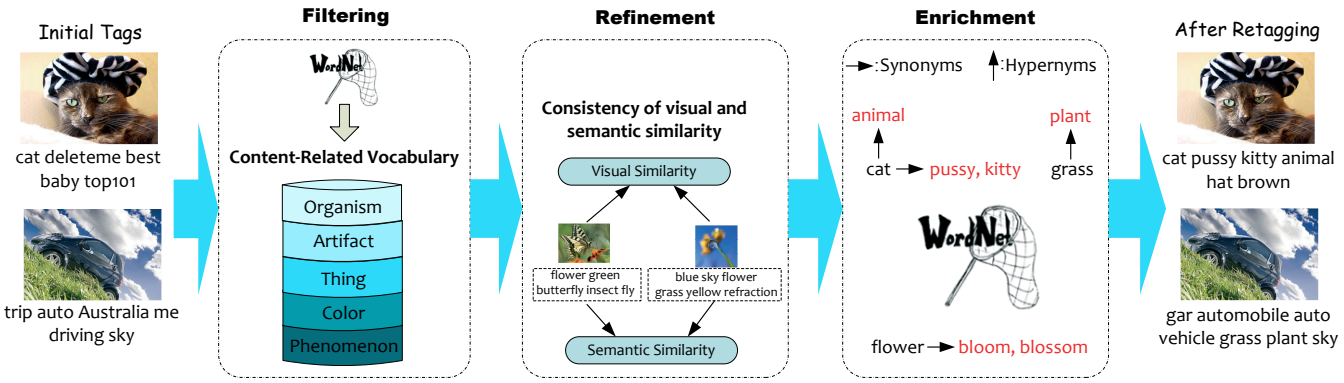

Figure 2: The schematic illustration of the image retagging approach. Tag filtering is first adopted to filter out content-unrelated tags. Then an optimization algorithm is performed to refine the tags. Finally, we augment each tag with its synonyms and hypernyms.

Sigurbjörnsson et al. [12] have provided the insights on how users tag their photos and what type of tags they are providing. Kennedy et al. [13] have evaluated the performance of the classifiers trained with Flickr images and the associated tags, and demonstrated that tags provided by Flickr users actually contain many noises. Yan et al. [14] have proposed a model that is able to predict the time cost of manual image tagging. Liu el al. [8] have proposed methods to analyze the relevance scores of tags with respect to the image. Weinberger et al. [15] have proposed a method to analyze the ambiguity of tags. Different tag recommendation methods have also been proposed that aim at helping users tag images more efficiently [12]. Despite these works have shown encouraging results, we can see that most of them focus on helping users tag images in more efficient ways or directly utilizing the tags as a knowledge source, whereas there is still a severe lack regarding improving the quality of tags.

The most related works to this paper are [16, 17], which showed some brief and preliminary formulations based on a similar motivation, but the scheme to be introduced in this paper will provide a much more comprehensive framework and much deeper theoretical analysis on the performance. Another related work to our retagging scheme is image annotation refinement [18, 19], which tries to improve the quality of inaccurate concept annotation results. As a pioneering work, Jin et al. [18] have used WordNet [7] to estimate the semantic correlation between to-be-annotated concepts and then highly-correlated concepts are preserved and weakly-correlated concepts are removed. However, this method has not taken the visual information of images into account, and it thus achieves only limited success. The effort in [19] has further leveraged visual clue, but it is still merely used to estimate the correlation between to-be-annotated concepts in order to perform belief propagation among concepts, and the usage is obviously not sufficient. This is reasonable for the annotation refinement methods, since typically visual information is already used in annotation algorithms, but in retagging the visual clue will be the most important information source to improve the tag quality. In our scheme, the tag refinement component will simultaneously model the visual and semantic clues. Experiments will show that this method is superior to the existing annotation refinement algorithms.

3. TAG FILTERING

As introduced in Section 1, many tags provided by users are actually intrinsically content unrelated, such as emotional tags and signaling tags (Signaling tags are those that indicate to other users to take action on images or recognize their importance, such as "delete me", "click me" and "best" [20]). Handling such tags is beyond the capability of computer vision and machine learning techniques. In addition, their existence may bring significant noises in analysis or learning algorithms, such as the tag refinement introduced in the next section. Therefore, we propose to filter out those content unrelated tags to ensure the performance of content-related tags can be significantly improved by the refining process.

To differentiate content-related tags from content-unrelated ones in a large tag set is not a trivial task. The most straightforward way is to check each word in an authoritative dictionary (such as *Merriam-Webster* dictionary) and ask humans to decide whether it is content-related based on their knowledge and understanding. However, it needs intensive labor costs. Another possible approach is to mine the relationship between tags and the content of images from large-scale data, such as the works in [21, 22]. However, a limitation of this approach is that it needs to be accomplished based on a set of low-level features. There is always a gap between these features and the actual image content. For example, the tags "red" and "green" are of course content-related, but these approaches may decide them as content-unrelated if SIFT features are applied. In our work, we solve this difficulty by taking advantage of lexical and domain knowledge, which is analogous to the approach in [23].

First, from the part-of-speech of words, we only consider nouns. For example, many photos on Flickr are tagged with verb "delete", adjective "beautiful", adverb "heavily" or numerals such as "2", "131", and "1001". Actually it is hard to say the non-noun words are definitely not related to image content, since there also may be several visual patterns for these words. For example, "beautiful" may relate to flower, scene or woman with high probability. But these non-noun words will be difficult to be analyzed or learned. Thus we restrict ourselves to the noun set of WordNet lexicon [7], which contains 114, 648 noun entries and the tags that are out of this set will be out of the scope of this paper.

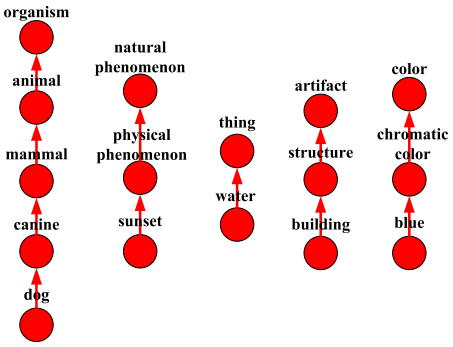

Figure 3: The decision process of some exemplary tags. Each tag successfully matches one of our pre-defined visual categories and thus is decided as content-related.

Then we further observe that noun tags are not always content-related. Some abstract nouns such as "science", "lifetime" and "distance", have weak connection with image content and thus also need to be filtered out. To this end, we adopt an automatic process to detect visual property of the tags as follows.

- We first empirically select a set of high level categories as a taxonomy of our domain knowledge in vision field. In this work, we choose "organism", "artifact", "thing", "color" and "natural phenomenon", since they cover a substantial part of tags and contain many frequently used concepts in computer vision and multimedia domains.
- For each entry in WordNet, there are one or more senses (a sense can be viewed as a meaning or explanation of the entry), in which WordNet provides a set of synonyms and hypernyms [24]. For each noun entry in WordNet lexicon, we traverse along the path that is composed of hypernyms of the given word until one of the pre-defined visual categories is matched. If the match succeeds, the word is decided as content-related, and otherwise it is decided as content-unrelated.

Figure 3 illustrates the decision process of several exemplary tags. In this way, we obtain a content-related vocabulary that contains 57,623 noun entries. Table 1 illustrates the number of entries for each category.

Visual Category	Number of entries	Percentage
Color	423	0.73%
Thing	4,484	7.78%
Artifact	14,087	24.45%
Organism	37,672	65.38%
Natural pheno.	957	1.66%

Table 1: Number of entries for each visual category in our content-related vocabulary.

It is worth mentioning that although our approach involves several heuristics, it can achieve fairly consistent results with humans (the detailed results will be demonstrated in Section 6). Our method is also with high flexibility. For example, the Wordnet lexical knowledge base can be easily replaced with other lexical base such as Wikipedia. Besides, we can easily further add categories to cover more entries. The vocabulary can also be manually refined by adding or removing entries. In addition, the number of content-related tags is still very large (57,623 entries) and we believe the description capability of this tag set is still high enough to describe most of the entities in general images. On the other hand, it is also feasible to extend our approach to non-noun tags like adjectives which appear frequently in the general user's tagging vocabulary by defining proper high-level semantic theme abstractions.

4. TAG REFINEMENT

In this section we introduce the tag refinement algorithm. We first present the optimization framework and then introduce an iterative bound optimization algorithm to solve it.

4.1 Tag Refinement Framework

Denote by $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ a social image collection, where n is the size of the image set. All unique tags appearing in this collection are gathered in a set $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}$, where m denotes the total number of unique tags. The initial tag membership for the whole image collection can be presented in a binary matrix $\widehat{\mathbf{Y}} \in \{0,1\}^{n \times m}$ whose element $\widehat{Y}_{ij}$ indicates the membership of tag t_j with respect to image x_i (i.e., if t_j is associated with image x_i, then $\widehat{Y}_{ij} = 1$, and otherwise $\widehat{Y}_{ij} = 0$). To represent the refinement results, we define another matrix $\mathbf{Y}$ whose element $Y_{ij} \geq 0$ denotes the confidence score of assigning tag t_j to image x_i. Denote by $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{im})^\top$ the confidence score vector of assigning the tags to the i-th image.

The tag refinement is accomplished based on two assumptions. The first is the consistency of visual and semantic similarities between images, i.e., the tags of two visually close images are expected to be similar[5], and the second is that the user-provided tags are relevant with high probability. We then introduce a regularization framework which contains two terms that indicate these two assumptions. Note that these assumptions may not hold for all images, but they are reasonable in most cases and the regularization scheme is able to automatically deal with several outliers.

We first compute the visual similarity between images. Let $\mathbf{W}$ denote a similarity matrix whose element W_{ij} indicates the visual similarity between images x_i and x_j, which can be directly computed based on a Gaussian function with a radius parameter σ, i.e.,

$$W_{ij} = exp(-\frac{\| x_i - x_j \|_2^2}{\sigma^2}), \quad (1)$$

where x_i and x_j denote the low level features of two images and $\| \cdot \|_2$ denotes the ℓ_2-norm.

The semantic similarity of two images is defined based on their tag sets. The most simple approach is to define it as the overlap between the two tag sets, i.e., $\mathbf{y}_i^\top \mathbf{y}_j$. However, this approach actually treats all tags independently without considering their correlation. For example, it will give zero

[5]It is worth noting that the assumption of the consistency between visual and semantic similarities might not hold for any two images, especially when the images of the same class (e.g. cars) have diversified visual features. However, it is always possible to find visually similar images with similar semantic theme in a large image collection, which satisfies our consistency assumption.

similarity for two images that do not share any common tag, but the images can still be very semantically similar if their tags are strongly correlated. To leverage the correlation of tags, we introduce the tag similarity matrix $\mathbf{S}$, in which the element $S_{ij} \geq 0$ indicates the tag similarity between tags t_i and t_j. In this work, we adopt Lin's similarity measure [25], a knowledge-based similarity metric which utilizes WordNet as knowledge base. It estimates the semantic similarity between t_i and t_j as

$$S_{ij} = \frac{2 * IC(lcs(t_i, t_j))}{IC(t_i) + IC(t_j)}, \tag{2}$$

where $IC(t_i)$ and $IC(t_j)$ are the information content[6] of t_i and t_j respectively, and $lcs(t_i, t_j)$ is their least common subsumer (LCS) in the WordNet taxonomy which represents the common ancestor of these two entries that has the maximum information content. But it is worth noting that our method is flexible, and the semantic similarities can be computed through other approaches as well, such as using Google distance [26]. Then we define the semantic similarity of images by a weighted dot product, i.e., $\mathbf{y}_i^\top \mathbf{S}\mathbf{y}_j = \sum_{k,l=1}^m Y_{ik} S_{kl} Y_{jl}$.

According to our consistency assumption, the visual similarity is expected to be close to semantic similarity, i.e., $W_{ij} \approx \mathbf{y}_i^\top \mathbf{S}\mathbf{y}_j$. This leads to the following formulation:

$$\begin{aligned} \min_{\mathbf{Y}} \quad & \sum_{i,j=1}^{n} (W_{ij} - \sum_{k,l=1}^{m} Y_{ik} S_{kl} Y_{jl})^2, \\ s.t. \quad & Y_{jl} \geq 0, \quad j = 1, 2, \ldots, n, \quad l = 1, 2, \ldots, m. \end{aligned} \tag{3}$$

We then consider the assumption that user-provided tags are relevant with high probability. Here we introduce the minimization of $\sum_{j=1}^n \sum_{l=1}^m (Y_{jl} - \widehat{Y}_{jl})^2 \exp(\widehat{Y}_{jl})$, where the term $\exp(\widehat{Y}_{jl})$ is a weighting constant for the initial user-provided tag assignment. A problem with the above formula is that Y_{ij} and $\widehat{Y}_{ij}$ may lie at different scales (from Eq. (3) we can see that Y_{ij} may be much smaller than 1, whereas the value of the initial tag membership $\widehat{Y}_{ij}$ is restricted to 0 or 1). To deal with this problem, we introduce a scaling factor α_j for each image and thus the optimization term turns to

$$\sum_{j=1}^{n} \sum_{l=1}^{m} (Y_{jl} - \alpha_j \widehat{Y}_{jl})^2 \exp(\widehat{Y}_{jl}). \tag{4}$$

Formally, we can summarize the above two assumptions into the following optimization problem:

$$\begin{aligned} \min_{\mathbf{Y}, \boldsymbol{\alpha}} \quad & \mathcal{L} = \sum_{i,j=1}^{n} (W_{ij} - \sum_{k,l=1}^{m} Y_{ik} S_{kl} Y_{jl})^2 \\ & + C \sum_{j=1}^{n} \sum_{l=1}^{m} (Y_{jl} - \alpha_j \widehat{Y}_{jl})^2 \exp(\widehat{Y}_{jl}), \\ s.t. \quad & Y_{jl}, \alpha_j \geq 0, \quad j = 1, 2, \ldots, n, \quad l = 1, 2, \ldots, m. \end{aligned} \tag{5}$$

where C is a weighting factor to modulate the two terms.

The above equation can be written in matrix form as

$$\begin{aligned} \min_{\mathbf{Y}, \mathbf{D}} \quad & \mathcal{L} = \| \mathbf{W} - \mathbf{Y}\mathbf{S}\mathbf{Y}^\top \|_F^2 + C \| (\mathbf{Y} - \mathbf{D}\hat{\mathbf{Y}}) \circ \mathbf{E} \|_F^2, \\ s.t. \quad & Y_{jl}, D_{jj} \geq 0, \end{aligned} \tag{6}$$

where $\circ$ indicates the pointwise product of matrices, the element E_{ij} of matrix $\mathbf{E}$ represents the weighting factor $\exp(\widehat{Y}_{jl})$ in Eq. (4), and $\mathbf{D}$ is an $n \times n$ diagonal matrix whose element $D_{jj} = \alpha_j$.

[6]The value of information content of a WordNet concept c is defined as $IC(c) = -\log(p(c))$, where $p(c)$ denotes the probability with which the concept c occurs in the text corpus.

4.2 Solution of the Optimization Problem

The coupling of $\mathbf{Y}$ and $\boldsymbol{\alpha}$ and the positive constraints of Y_{jl} and α_j make the above optimization framework difficult to solve. In this work, we propose an efficient iterative bound optimization method to obtain its solution, which is analogous to [27]. The basic idea is to first bound the objective function with a computationally feasible function, and then implement an iterative algorithm to obtain its optimal solution. The following two theorems provide the upper bound of $\mathcal{L}$ and its solution.

THEOREM 1. *The objective function $\mathcal{L}$ in Eq. (5) has an upper bound $\mathcal{L}'$ as follows.*

$$\begin{aligned} \mathcal{L} \leq \mathcal{L}' = & \sum_{i,j=1}^{n} \Big(W_{ij}^2 + \sum_{l=1}^{m} [\widetilde{\boldsymbol{Y}}\boldsymbol{S}\widetilde{\boldsymbol{Y}}^\top]_{ij} [\widetilde{\boldsymbol{Y}}\boldsymbol{S}]_{il} \frac{Y_{jl}^4}{\widetilde{Y}_{jl}^3} \\ & - 4 \sum_{l=1}^{m} W_{ij} [\widetilde{\boldsymbol{Y}}\boldsymbol{S}]_{il} \widetilde{Y}_{jl} \log Y_{jl} - 2 W_{ij} [\widetilde{\boldsymbol{Y}}\boldsymbol{S}\widetilde{\boldsymbol{Y}}^\top]_{ij} \\ & + 4 \sum_{k=1}^{m} W_{ij} [\boldsymbol{S}\widetilde{\boldsymbol{Y}}^\top]_{kj} \log \widetilde{Y}_{ik} \Big) \\ & + C \sum_{j=1}^{n} \sum_{l=1}^{m} \Big(Y_{jl}^2 - 2\alpha_j \widehat{Y}_{jl} \widetilde{Y}_{jl} (\log \frac{Y_{jl}}{\widetilde{Y}_{jl}} + 1) \\ & + \alpha_j^2 \widehat{Y}_{jl}^2 \Big) \exp(\widehat{Y}_{jl}), \end{aligned} \tag{7}$$

where $\tilde{\boldsymbol{Y}}$ can be any $n \times m$ matrix with positive entries.

THEOREM 2. *The optimal solution of the upper bound $\mathcal{L}'$ is given by*

$$\begin{cases} Y_{jl} = \left[\frac{-C \exp(\hat{Y}_{jl}) \tilde{Y}_{jl}^3 + \sqrt{M}}{4 [\tilde{\boldsymbol{Y}}\boldsymbol{S}\tilde{\boldsymbol{Y}}^\top \tilde{\boldsymbol{Y}}\boldsymbol{S}]_{jl}} \right]^{\frac{1}{2}}, \\ \alpha_j = \frac{\sum_{l=1}^m \tilde{Y}_{jl} (\log Y_{jl} - \log \tilde{Y}_{jl} + 1)}{\sum_{l=1}^m \tilde{Y}_{jl}}, \end{cases} \tag{8}$$

where $M = (C \exp(\hat{Y}_{jl}))^2 + 8 U_{jl} \tilde{Y}_{jl}^4 (2 [\boldsymbol{W}\tilde{\boldsymbol{Y}}\boldsymbol{S}]_{jl} + C \alpha_j \hat{Y}_{jl} \exp(\hat{Y}_{jl}))$ *with* $U_{jl} = [\tilde{\boldsymbol{Y}}\boldsymbol{S}\tilde{\boldsymbol{Y}}^\top \tilde{\boldsymbol{Y}}\boldsymbol{S}]_{jl}$.

Both of the two theorems can be verified with algebraic manipulations, and their proofs are provided in Appendices A and B, respectively. With the obtained solution in Eq. (8), we present an iterative algorithm to approximate the optimal solution of the upper bound function $\mathcal{L}'$, as shown in Algorithm 1.

Algorithm 1 Iterative Optimization Algorithm

Input: Visual similarity matrix $\mathbf{W}$, Semantic similarity matrix $\mathbf{S}$, Weighting parameter C.
Output: $\mathbf{Y}$,$\boldsymbol{\alpha}$.

1. Randomly initialize $\mathbf{Y}$ and $\boldsymbol{\alpha}$ that satisfy the constraints in Eq. (5). Their initial values will not influence the final results.
2. Repeat the following steps until convergence:
   - Fix $\boldsymbol{\alpha}$, update $\mathbf{Y}$ using the upper equation in Eq. (8).
   - Fix $\mathbf{Y}$, update $\boldsymbol{\alpha}$ using the lower equation in Eq. (8).

In fact the procedure above not only minimizes the upper bound function $\mathcal{L}'$ but also obtains the optimal solution of the objective function $\mathcal{L}$, i.e., we have the following theorem:

Figure 4: An example image on Flickr and its associated tags. The synonyms and hypernyms are not simultaneously provided by the user.

THEOREM 3. *The process in Algorithm 1 will converge and leads to the minimum of the objective function $\mathcal{L}$.*

The proof of this theorem can be found in Appendix C. With the computed confidence score matrix $\mathbf{Y}$, we can easily generate the refined tags of an image x_i by setting a threshold for the confidence scores Y_{ij}.

5. FURTHER ENRICHMENT

Although the tag refinement approach can effectively improve the quality of tags, it is still unable to introduce new tags for an image if these tags do not exist in the original image collection. However, this difficulty can be alleviated in certain extent through adding synonyms and hypernyms. Actually, the synonym and hypernym are both valuable indexing keywords that can be utilized to describe the visual content. As an example, suppose we have an image annotated with the tag "kitty". We can observe that the synonyms and hypernyms of "kitty", such as "kitten", "pussy", "feline" and "animal" are also relevant to the image. However, most users will not spend time to enter these tags. Here is a simple empirical validation. We randomly select 1,000 images that are tagged with "kitty" on Flickr and then check the synonyms and hypernyms. Table 2 illustrates several statistical results, including the percentages of images that are also tagged with its synonyms ("kitten" or "pussy") or hypernyms ("feline" or "animal"). As can be seen, only a small percentage of images are simultaneously tagged with "kitty" and its synonyms or hypernyms.

Relation with "kitty"	Exists with "kitty" together	Percentage of images
Synonym	*kitten*	30.6%
	pussy	2.3%
Hypernym	*animal*	24.8%
	feline	28.0%

Table 2: Percentage of images that are simultaneously tagged with "kitty" and its synonyms ("kitten", "pussy") as well as hypernyms ("feline", "animal").

The missing of such tags will degrade the performance of tag-based applications. For example, when users perform search with query "kitten" or "animal", the photo in Figure 4 will not be returned. To address this issue, we adopt a tag enrichment step based on lexical knowledge to expand the synonyms and hypernyms of the refined tags, and this will be an effective strategy to improve the coverage of the tags.

We take advantage of the WordNet lexicon again. As discussed previously, the tags after filtering and refinement are all content-related nouns, and each of them has a coordinate in WordNet. Hence, we apply the lemma entries in WordNet lexicon as the potential enrichment candidate words. Take tag "kitty" for example, it matches the visual category "organism" in its 4-th sense, and thus we collect all the synonyms and hypernyms in this sense, as illustrated in Figure 5.

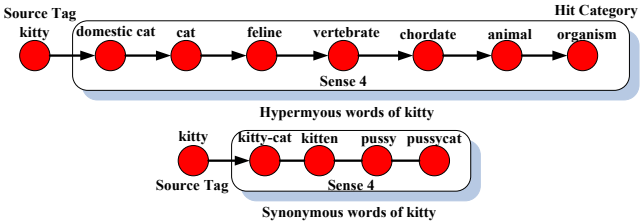

Figure 5: The hypernyms and synonyms of tag "kitty" can be collected based on the WordNet lexicon.

6. EMPIRICAL STUDY

We implement and evaluate our retagging approach on a collection of social images. As the scheme consists of three components, we carry out experiments to test the performance for each of them.

6.1 Data Set

We evaluate our approach on social images that are collected from Flickr. We select ten most popular queries, including *cat*, *automobile*, *mountain*, *water*, *sea*, *bird*, *tree*, *sunset*, *flower* and *sky*, and then perform tag-based image search with "ranking by interestingness" option. The top 5,000 returned images for each query are collected together with their associated information, including tags, uploading time, user identifier, etc. In this way, we obtain a social image collection consisting of 50,000 images and 106,565 unique tags.

For each image, we extract 428-dimensional features, including 225-dimensional block-wise color moment features generated from 5-by-5 fixed partition of the image, 128-dimensional wavelet texture features, and 75-dimensional edge distribution histogram features. These low-level feature vectors are then used for the calculation of visual similarity between the images with Eq. (1).

6.2 Evaluation of Tag Filtering

As introduced in Section 3, we have constructed a content-related vocabulary which is a subset of the WordNet lexicon. We first match each of the 106,565 tags in our social image collection with the noun entries in WordNet. In this way, 11,861 noun tags are obtained, occupying only 11% of all tags in our collection. We then further match these tags with the entries in the content-related vocabulary that includes 5 high level categories, and thus obtain 4,556 tags. Table 3 illustrates the detailed results.

Now we conduct a simple testing to evaluate this strategy. We randomly select 500 tags from the 11,861 noun tags and then ask human assessors to label each tag as "content-related" or "content-unrelated". We invite five researchers

Categories		Tag number	Percentage
Content-Unrelated		7,305	61.6%
Content-Related	*Color*	83	0.70%
	Thing	483	4.07%
	Artifact	2,511	21.2%
	Organism	1,259	10.6%
	Natural Pheno.	220	1.85%

Table 3: Statistics of our tag detection results. There are 4,556 content-related tags in all in our social image collection.

in multimedia research community to perform the labeling task based on their knowledge. The voting results are considered as the ground truths of these tags. We regard our method as a binary classification process, and employ accuracy measurement as evaluation metric. Table 4 illustrates the results, from which we can see that the accuracy measurement is as high as 0.83. Considering the simplicity and the computational efficiency of our method, the result is satisfactory[7].

	Related	Unrelated
Related	132	58
Unrelated	27	283
Accuracy	0.83	

Table 4: Confusion matrix obtained by our content-related tag detection method.

6.3 Evaluation of Tag Refinement

The proposed tag refinement method is also evaluated on the Flickr dataset. To reduce computational cost, we first perform the K-means clustering algorithm to divide the 50,000 images into 20 groups. Then the tag refinement algorithm is implemented for each group separately. As introduced in Section 4, our tag refinement algorithm is developed mainly based on the similarities of images. Since most similar images will be clustered into the same groups, the clustering strategy will not degrade the performance of refinement while being able to significantly reduce computational costs. The radius parameter σ in Eq. (1) is set to the median value of all the pair-wise Euclidean distances between images in the same cluster, and the parameter C in Eq. (5) is empirically set to 10. We compare the following three methods:

- Baseline, i.e., keeping the original tags.
- Content Based Annotation Refinement (CBAR). We adopt the method proposed in [19].
- Our tag refinement method.

Note that actually CBAR and our tag refinement method both produce confidence scores for tags. Therefore, for these two methods we rank the tags of each image based on their confidence scores and then keep the top m tags where m is the number of the original tags. The average number of tags per image is about 5. The ground truths of the tags are voted by three volunteers. If a tag is relevant to the image, it is labeled as positive, and otherwise it is negative. However, manually labeling all the image-tag pairs will be too labor-intensive, and thus here we randomly select 2,500 images as the evaluation set. We adopt precision/recall/F1-measure as performance measurements. But a problem is that the estimation of recall measurement needs to know the full set of relevant tags for each image, but in our case it is unknown. So here we adopt an alternative strategy. We gather the tags obtained by CBAR and our method as well as the original tags for each image, and then the positive tags among them are regarded as the full relevant set.

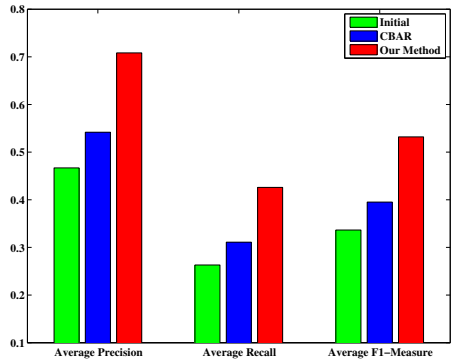

Figure 6: Performance comparison of the baseline and the two tag refinement methods.

Figure 6 shows the precision, recall and F1-measure measurements obtained by the three methods, averaged over all evaluation images. We can see that CBAR and our method both outperform the baseline method, but the improvement of CBAR is limited. As analyzed in Section 2, this is due to the fact that visual information has not been sufficiently explored in CBAR. Our tag refinement method performs much better than the baseline and the CBAR methods.

We also study the performance of keeping varied number of tags. Here we only compare CBAR and our tag refinement method, since there is no ranking in the original tags. The results are illustrated in Figure 7. From the figure we can clearly see that our method consistently outperforms the CBAR method.

6.4 Evaluation of Tag Enrichment

In this section, we show the results of tag enrichment. As introduced in Section 5, we obtain the synonyms and hypernyms of each tag based on the WordNet lexicon, and then a tag-based search is performed on Flickr website to obtain their usage frequencies to decide whether to enrich them. We empirically set the threshold to 10,000 in this work, i.e., the tags with usage frequencies under 10,000 will not be enriched.

We still utilize the 2,500 evaluation images to validate performance of tag enrichment. The ground truths of the enriched tags are also decided by the three labelers. The performance measurements before and after enrichment are illustrated in Table 5 (the recall measurement is computed

[7]Recently Overell et al. [23] have conducted a study on classifying tags into 25 semantic categories. They have leveraged the knowledge sources of both Wikipedia and WordNet and used a SVM classifier. The study demonstrates that this approach significantly outperforms other methods, but the classification error rate is still as high as 0.38. This result indicates that tag classification is actually a challenging task.

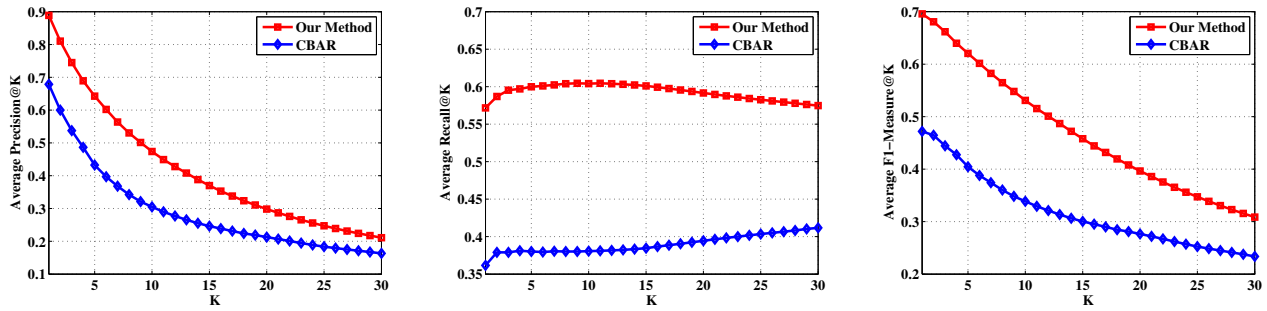

Figure 7: Performance comparison of the CBAR and our methods when keeping different number of tags.

Method	Precision	Recall	F1-Measure	Relevant tag number
Before Enrichment	0.71	0.34	0.46	3.09 (4.80 in all)
After Enrichment	0.90	0.66	0.76	9.34 (10.38 in all)

Table 5: Quality comparison of the tags before and after enrichment. Each measurement has been averaged over all evaluation images.

Methods	Using images with original tags	Using retagged images
Precision in average	0.221	0.340

Table 6: Performance comparison of annotation before and after performing retagging.

in the same way as in the last section, i.e., collect all the relevant tags obtained by different methods and regard them as the full set). From the table we can see that all performance evaluation measurements are improved after enrichment. We can see that the recall measurement is improved from 0.34 to 0.66, and this indicates the coverage of the tags is significantly improved. Actually our experimental results show that the average number of relevant tags for an image is improved from 3.09 to 9.34 after enrichment.

7. APPLICATION

It is widely acknowledged that user-generated tags are valuable metadata that can be used in many applications. In this section, we introduce two potential application scenarios that could benefit from image retagging.

- Tag-based image search. The images that contain the query tag are returned in descending order in terms of their confidence scores.
- Tag-based automatic annotation. For a given image, we provide the tags of its visual neighbors as the potential labels of the image.

7.1 Tag-Based Image Search

Since the retagging scheme improves the quality of tags, tag-based image search can be naturally improved. In addition, our retagging scheme also brings a ranking scheme for tag-based image search since it not only adds and removes tags but also assigns them confidence scores. Currently online image sharing websites support tag-based image search and browsing, but they lack an effective scheme to rank the search result according to the relevance levels of the images. Take Flickr for example, it only provides two options to rank tag-based image search results. One is "most recent", which ranks the most recently uploaded images on top, and the other is "most interesting" which ranks images by "interestingness", a measure that takes clickthrough, comments and so on into consideration. These two schemes do not take relevance into account. As introduced in Section 4, our retagging scheme can obtain the confidence scores of tags with respect to their associated images, and we can regard them as an indication of relevance levels and provide a relevance-based ranking.

To demonstrate the effectiveness of this approach, we conduct experiment on the 50,000 Flickr images and perform tag-based image search with ten tags, i.e., *cat*, *automobile*, *flower*, *sea*, *bird*, *sunset*, *mountain*, *water*, *sky* and *tree*. For each ranking list, the images are decided as "relevant" or "irrelevant" with respect to the query tag by human labelers, and we employ Average Precision (AP) as our evaluation metric. Give a ranked list L with length n, AP is defined as

$$AP = \frac{1}{R}\sum_{j=1}^{n}\frac{R_j}{j}I_j, \tag{9}$$

where R is the number of relevant instances in L, R_j the number of relevant instances among the top j instances, $I_j = 1$ if the j-th instance is relevant and 0 otherwise. To evaluate the overall performance, we use Mean AP, the mean value of AP scores over all queries.

We compare our approach with the two schemes provided by Flickr: (1) interestingness-based ranking; and (2) uploading time-based ranking. Figure 8 illustrates the mean AP measurement at different return depths and we can see that our approach consistently outperforms the other two ranking methods. This demonstrates our approach can remarkably improve tag-based image search in terms of relevance.

7.2 Tag-Based Automatic Annotation

Leveraging tagged images as training data for annotation is an important application. Traditional image annotation algorithms frequently suffer from the insufficiency of training

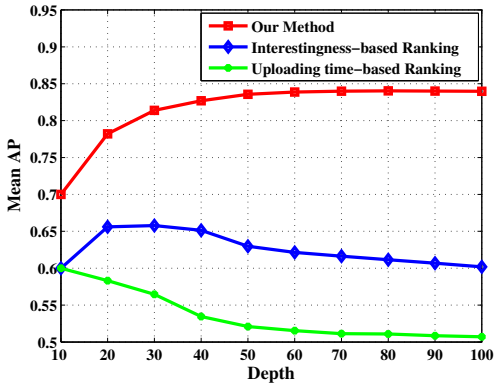

Figure 8: The performance comparison of different ranking methods for tag-based image search.

data. But actually nowadays we can easily collect enormous tagged images (such as from Flickr), and these images can be used as labeled data to help annotation, such as the work in [28]. Here we show that this application can benefit a lot from our retagging scheme, since the quality of image tags has been significantly improved.

The experimental setting is as follows. We randomly select 500 images from a personal photo album as a testing set. Then for each testing image, we perform the k-NN method to decide its labels. More specifically, we find its k-nearest neighbors in our tagged Flickr data (50,000 in all) and then collect their tags. The confidence of each tag is estimated as its appearing frequency in the neighborhood. We then regard the T most confident tags as the labels of the image. In our experiments, the parameters K and T are both empirically set to 10.

The experimental results are illustrated in Table 6. We can see that performing retagging on the training data can help improve the precision measurement from 0.221 to 0.340, i.e., there are 2.2 true labels for a testing image in average if we use the original tags, and there will be 3.4 true labels if we perform retagging. The relative improvement is 53.8%. Therefore, our retagging method can help obtain more accurate tag-based annotation results.

8. CONCLUSION

We have introduced an image retagging scheme that aims at improving the quality of the tags associated with social images in terms of content relevance. The scheme consists of three components, i.e., tag filtering, tag refinement and further enrichment. Experiments on real-world social image dataset have demonstrated its effectiveness. We have also shown two applications that benefit from the retagging scheme. Although we have put more emphasis on Flickr in this work, the proposed retagging framework is flexible and can be easily extended to deal with a variety of online media repositories, such as Zooomr as well as any other media databases with noisy and incomplete tags.

9. REFERENCES

[1] P. Anderson. What is web 2.0? Ideas, technologies and implications for education. *JISC Technical Report*, 2007.

[2] M. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *TOMCCAP*, 2(1):1-19, 2006.

[3] S. Golder and B. Huberman. Usage patterns of collaborative tagging systems. *JIS*, 32(2):198-208, 2006.

[4] K. Matusiak. Towards user-centered indexing in digital image collections. *OCLC Systems and Service*, 22(4):283-298, 2006.

[5] J. Li and J. Wang. Real-time computerized annotation of pictures. *TPAMI*, 30(6):985-1002, 2008.

[6] X.-S. Hua and G. Qi. Online multi-label active annotation: Towards large-scale content-based video search. In *MM*, pages 141-150, 2008.

[7] C. Fellbaum. Wordnet: An electronic lexical database. *Bradford Books*, 1998.

[8] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag ranking. In *WWW*, pages 351-360, 2009.

[9] D. Liu, X.-S. Hua, M. Wang, and H.-J. Zhang. Boost search relevance for tag-based social image retrieval. In *ICME*, pages 1636-1639, 2009.

[10] Z.-J. Zha, L. Yang, T. Mei, M. Wang, and Z. Wang. Visual query suggestion. In *MM*, pages 15-24, 2009.

[11] M. Ames and M. Naaman. Why we tag: Motivations for annotation in mobile and online media. In *CHI*, pages 971-980, 2007.

[12] B. Sigurbjörnsson and R. Zwol. Flickr tag recommendation based on collective knowledge. In *WWW*, pages 327-336, 2008.

[13] L. Kennedy, S.-F. Chang, and I. Kozintsev. To search or to label? Predicting the performance of search-based automatic image classifiers. In *MIR*, pages 249-258, 2006.

[14] R. Yan, A. Natsev, and M. Campbell. A learning-based hybrid tagging and browsing approach for efficient manual image annotation. In *CVPR*, pages 1-8, 2008.

[15] K. Weinberger, M. Slaney, and R. Zwol. Resolving tag ambiguity. In *MM*, pages 111-120, 2008.

[16] D. Liu, X.-S. Hua, M. Wang, and H.-J. Zhang. Retagging social images based on visual and semantic consistency. In *WWW*, pages 1149-1150, 2010.

[17] D. Liu, M. Wang, J. Yang, X.-S. Hua, and H.-J. Zhang. Tag quality improvement for social images. In *ICME*, pages 350-353, 2009.

[18] Y. Jin, L. Khan, L. Wang, and M. Awad. Image annotation by combining multiple evidence & wordNet. In *MM*, pages 706-715, 2005.

[19] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. Content-based image annotation refinement. In *CVPR*, pages 1-8, 2007.

[20] B. Dennis. Foragr: Collaboratively tagged photographs and social information visualization. In *WWW*, 2006.

[21] Y. Lu, L. Zhang, Q. Tian, and W. Ma. What are the high-level concepts with small semantic gaps? In *CVPR*, pages 1-8, 2008.

[22] K. Yanai and K. Barnard. Image region entropy: A measure of "visualness" of web images associated with one concept. In *MM*, pages 419-422, 2005.

[23] S. Overell, B. Sigurbjörnsson, and R. Zwol. Classifying tags uing open content resources. In *WSDM*, pages 64-73, 2009.

[24] A. Torralba, R. Fergus, and W. Freeman. 80 million tiny images: A large dataset for non-parametric object and scene recognition. *TPAMI*, 30(11):1958-1970, 2008.

[25] D. Lin. Using syntatic dependency as local context to resolve word sense ambiguity. In *ACL*, pages 64-71, 1997.

[26] R. Cilibrasi and P. Vitanyi. The google similarity distance. *TKDE*, 19(3):370-383, 2007.

[27] Y. Liu, R. Jin, and L. Yang. Semi-supervised multi-label learning by constrained non-negative matrix factorization. In *AAAI*, pages 421-426, 2006.

[28] S.-F. Chang, J. He, Y. Jiang, E. Khoury, C. Ngo, A. Yanagawa, and E. Zavesky. Columbia University/VIREO-CityU/IRIT TRECVID2008 high-level

feature extraction and interactive video search. In *NIST TRECVID Workshop*, 2008.

[29] D. Lee and H. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556-562, 2001.

APPENDIX

A. PROOF OF THEOREM 1

PROOF. First, we derive the upper bound of the first term of the objective function in Eq. (5) as follows

$$\begin{aligned}
&\textstyle\sum_{i,j=1}^{n}(W_{ij}-\sum_{k,l=1}^{m}Y_{ik}S_{kl}Y_{jl})^2\\
&=\textstyle\sum_{i,j=1}^{n}(W_{ij}-\frac{[\tilde{\mathbf{Y}}\mathbf{S}\tilde{\mathbf{Y}}^\top]_{ij}}{[\tilde{\mathbf{Y}}\mathbf{S}\tilde{\mathbf{Y}}^\top]_{ij}}\sum_{k,l=1}^{m}\frac{\tilde{Y}_{ik}S_{kl}\tilde{Y}_{jl}}{\tilde{Y}_{ik}S_{kl}\tilde{Y}_{jl}}Y_{ik}S_{kl}Y_{jl})^2\\
&\le\textstyle\sum_{i,j=1}^{n}\Big(\sum_{k,l=1}^{m}\frac{\tilde{Y}_{ik}S_{kl}\tilde{Y}_{jl}}{[\tilde{\mathbf{Y}}\mathbf{S}\tilde{\mathbf{Y}}^\top]_{ij}}(W_{ij}-[\tilde{\mathbf{Y}}\mathbf{S}\tilde{\mathbf{Y}}^\top]_{ij}\frac{Y_{ik}S_{kl}Y_{jl}}{\tilde{Y}_{ik}S_{kl}\tilde{Y}_{jl}})^2\Big)\\
&=\textstyle\sum_{i,j=1}^{n}\Big(W_{ij}^2+\sum_{k,l=1}^{m}(\frac{[\tilde{\mathbf{Y}}\mathbf{S}\tilde{\mathbf{Y}}^\top]_{ij}}{\tilde{Y}_{ik}S_{kl}\tilde{Y}_{jl}}Y_{ik}^2S_{kl}^2Y_{jl}^2-2W_{ij}Y_{ik}S_{kl}Y_{jl})\Big)\\
&\le\textstyle\sum_{i,j=1}^{n}\Big(W_{ij}^2+\frac{1}{2}\sum_{k,l=1}^{m}[\tilde{\mathbf{Y}}\mathbf{S}\tilde{\mathbf{Y}}^\top]_{ij}\tilde{Y}_{ik}S_{kl}\tilde{Y}_{jl}(\frac{Y_{ik}^4}{\tilde{Y}_{ik}^4}+\frac{Y_{jl}^4}{\tilde{Y}_{jl}^4})\\
&-2\textstyle\sum_{k,l=1}^{m}W_{ij}\tilde{Y}_{ik}S_{kl}\tilde{Y}_{jl}(1+\log Y_{ik}+\log Y_{jl}\\
&-\log\tilde{Y}_{ik}-\log\tilde{Y}_{jl})\Big)\\
&=\textstyle\sum_{i,j=1}^{n}\Big(W_{ij}^2+\sum_{l=1}^{m}[\tilde{\mathbf{Y}}\mathbf{S}\tilde{\mathbf{Y}}^\top]_{ij}[\tilde{\mathbf{Y}}\mathbf{S}]_{il}\frac{Y_{jl}^4}{\tilde{Y}_{jl}^3}\\
&-4\textstyle\sum_{l=1}^{m}W_{ij}[\tilde{\mathbf{Y}}\mathbf{S}]_{il}\tilde{Y}_{jl}\log Y_{jl}-2W_{ij}[\tilde{\mathbf{Y}}\mathbf{S}\tilde{\mathbf{Y}}^\top]_{ij}\\
&+4\textstyle\sum_{k=1}^{m}W_{ij}\tilde{Y}_{ik}[\mathbf{S}\tilde{\mathbf{Y}}^\top]_{kj}\log\tilde{Y}_{ik}\Big),
\end{aligned}\tag{10}$$

where $\tilde{\mathbf{Y}}$ can be any $n\times m$ matrix. The first inequality holds due to convexity of the square function, i.e., $(\sum_{i=1}^{m}x_i)^2\le(\sum_{i=1}^{m}\frac{1}{\lambda_i}x_i^2)(\sum_{i=1}^{m}\lambda_i)$ and the second inequality is derived from the property of the square function, i.e., $xy\le(x^2+y^2)/2$ as well as the concave property of logarithm function, i.e., $\log x\le x-1$.

We then derive the upper bound of the second term of the objective function in Eq. (5) based on the concaveness of logarithm function as follows

$$\begin{aligned}
&C\textstyle\sum_{j=1}^{n}\sum_{l=1}^{m}(Y_{jl}-\alpha_j\hat{Y}_{jl})^2\exp(\hat{Y}_{jl})\\
&=C\textstyle\sum_{j=1}^{n}\sum_{l=1}^{m}(Y_{jl}^2-2\alpha_j\hat{Y}_{jl}Y_{jl}+\alpha_j^2\hat{Y}_{jl}^2)\exp(\hat{Y}_{jl})\\
&\le C\textstyle\sum_{j=1}^{n}\sum_{l=1}^{m}\Big(Y_{jl}^2-2\alpha_j\hat{Y}_{jl}\tilde{Y}_{jl}(\log\frac{Y_{jl}}{\tilde{Y}_{jl}}+1)+\alpha_j^2\hat{Y}_{jl}^2\Big)\exp(\hat{Y}_{jl}).
\end{aligned}\tag{11}$$

Combining the upper bounds in Eq. (10) and Eq. (11), we have

$$\begin{aligned}
\mathcal{L}\le\mathcal{L}'=&\textstyle\sum_{i,j=1}^{n}\Big(W_{ij}^2+\sum_{l=1}^{m}[\tilde{\mathbf{Y}}\mathbf{S}\tilde{\mathbf{Y}}^\top]_{ij}[\tilde{\mathbf{Y}}\mathbf{S}]_{il}\frac{Y_{jl}^4}{\tilde{Y}_{jl}^3}\\
&-4\textstyle\sum_{l=1}^{m}W_{ij}[\tilde{\mathbf{Y}}\mathbf{S}]_{il}\tilde{Y}_{jl}\log Y_{jl}-2W_{ij}[\tilde{\mathbf{Y}}\mathbf{S}\tilde{\mathbf{Y}}^\top]_{ij}\\
&+4\textstyle\sum_{k=1}^{m}W_{ij}[\mathbf{S}\tilde{\mathbf{Y}}^\top]_{kj}\log\tilde{Y}_{ik}\Big)\\
&+C\textstyle\sum_{j=1}^{n}\sum_{l=1}^{m}\Big(Y_{jl}^2-2\alpha_j\hat{Y}_{jl}\tilde{Y}_{jl}(\log\frac{Y_{jl}}{\tilde{Y}_{jl}}+1)\\
&+\alpha_j^2\hat{Y}_{jl}^2\Big)\exp(\hat{Y}_{jl}),
\end{aligned}\tag{12}$$

which completes the proof. □

B. PROOF OF THEOREM 2

PROOF. Taking the derivative of $\mathcal{L}'$ with respect to Y_{jl} and α_j and setting them equal to zero, we can obtain the following equations:

$$\begin{cases}
4\sum_{i=1}^{n}[\tilde{\mathbf{Y}}\mathbf{S}\tilde{\mathbf{Y}}^\top]_{ij}[\tilde{\mathbf{Y}}\mathbf{S}]_{il}\frac{Y_{jl}^3}{\tilde{Y}_{jl}^3}-4\sum_{i=1}^{n}W_{ij}[\tilde{\mathbf{Y}}\mathbf{S}]_{il}\frac{\tilde{Y}_{jl}}{Y_{jl}}\\
+C(2Y_{jl}-2\alpha_j\hat{Y}_{jl}\frac{\tilde{Y}_{jl}}{Y_{jl}})\exp(\hat{Y}_{jl})=0,\\
\alpha_j\sum_{l=1}^{m}\hat{Y}_{jl}-\sum_{l=1}^{m}\tilde{Y}_{jl}(\log\frac{Y_{jl}}{\tilde{Y}_{jl}}+1)=0.
\end{cases}\tag{13}$$

With some algebraic manipulation, we can obtain the solution of the above equation set as

$$\begin{cases}
Y_{jl}=\left[\frac{-C\exp(\hat{Y}_{jl})\tilde{Y}_{jl}^3+\sqrt{M}}{4[\tilde{\mathbf{Y}}\mathbf{S}\tilde{\mathbf{Y}}^\top\tilde{\mathbf{Y}}\mathbf{S}]_{jl}}\right]^{\frac{1}{2}},\\
\alpha_j=\frac{\sum_{l=1}^{m}\tilde{Y}_{jl}(\log Y_{jl}-\log\tilde{Y}_{jl}+1)}{\sum_{l=1}^{m}\tilde{Y}_{jl}},
\end{cases}\tag{14}$$

where $M=(C\exp(\hat{Y}_{jl}))^2+8U_{jl}\tilde{Y}_{jl}^4(2[\mathbf{W}\tilde{\mathbf{Y}}S]_{jl}+C\alpha_j\hat{Y}_{jl}\exp(\hat{Y}_{jl}))$ and $U_{jl}=[\tilde{\mathbf{Y}}\mathbf{S}\tilde{\mathbf{Y}}^\top\tilde{\mathbf{Y}}\mathbf{S}]_{jl}$, and it completes the proof. □

C. PROOF OF THEOREM 3

To prove the convergence of the iterative algorithm in Algorithm 1, we need to first introduce an auxiliary function, which is analogous to [29].

Definition 1. (**Auxiliary Function**) *A function $G(\Theta,\Theta')$ is called an auxiliary function of function $F(\Theta)$ if*

$$G(\Theta,\Theta')\ge F(\Theta),G(\Theta,\Theta)=F(\Theta)\tag{15}$$

holds for any Θ, Θ'.

Let $\{\mathbf{Y}^t\}$ be the series of matrices obtained in different iterations of Algorithm 1. We can define

$$\mathbf{Y}^{t+1}=\arg\min_{\mathbf{Y}}G(\mathbf{Y},\mathbf{Y}^t),\tag{16}$$

where $G(\mathbf{Y},\mathbf{Y}^t)$ is the auxiliary function of our objective function $\mathcal{L}(\mathbf{Y})$ in Eq. (5). Then we have

$$\mathcal{L}(\mathbf{Y}^t)=G(\mathbf{Y}^t,\mathbf{Y}^t)\ge G(\mathbf{Y}^{t+1},\mathbf{Y}^t)\ge\mathcal{L}(\mathbf{Y}^{t+1}).\tag{17}$$

Thus we can see that $\mathcal{L}(\mathbf{Y}^t)$ is monotonically decreasing. Then we only need to find an appropriate auxiliary function $G(\mathbf{Y},\mathbf{Y}^t)$ and its global optimal. The following lemma presents such an auxiliary function for $\mathcal{L}(\mathbf{Y}^t)$.

LEMMA 1. *The following function*

$$\begin{aligned}
G(\boldsymbol{Y},\boldsymbol{Y}')=&\textstyle\sum_{i,j=1}^{n}\Big(W_{ij}^2+\sum_{l=1}^{m}[\boldsymbol{Y}'\boldsymbol{S}\boldsymbol{Y}'^\top]_{ij}[\boldsymbol{Y}'\boldsymbol{S}]_{il}\frac{Y_{jl}^4}{Y_{jl}'^3}\\
&-4\textstyle\sum_{l=1}^{m}W_{ij}[\boldsymbol{Y}'\boldsymbol{S}]_{il}Y_{jl}'\log Y_{jl}-2W_{ij}[\boldsymbol{Y}'\boldsymbol{S}\boldsymbol{Y}'^\top]_{ij}\\
&+4\textstyle\sum_{k=1}^{m}W_{ij}Y_{ik}'[\boldsymbol{S}\boldsymbol{Y}'^\top]\log Y_{ik}'\Big)\\
&+C\textstyle\sum_{j=1}^{n}\sum_{l=1}^{m}\Big(Y_{jl}^2-2\alpha_j\hat{Y}_{jl}Y_{jl}'(\log\frac{Y_{jl}}{Y_{jl}'}+1)+\alpha_j^2Y_{jl}'^2\Big)\exp(\hat{Y}_{jl})
\end{aligned}\tag{18}$$

is the auxiliary function of the objective function $\mathcal{L}(\boldsymbol{Y})$ in Eq. (5). Furthermore, it is convex with respect to $\boldsymbol{Y}$ and its global minimum is

$$Y_{jl}=\left[\frac{-C\exp(\hat{Y}_{jl})Y_{jl}'^3+\sqrt{M}}{4[\boldsymbol{Y}'\boldsymbol{S}\boldsymbol{Y}'^\top\boldsymbol{Y}'\boldsymbol{S}]_{jl}}\right]^{\frac{1}{2}},\tag{19}$$

where $M=(C\exp(\hat{Y}_{jl}))^2+8U_{jl}Y_{jl}'^4(2[\boldsymbol{W}\boldsymbol{Y}'\boldsymbol{S}]_{jl}+C\alpha_j\hat{Y}_{jl}\exp(\hat{Y}_{jl}))$ and $U_{jl}=[\boldsymbol{Y}'\boldsymbol{S}\boldsymbol{Y}'^\top\boldsymbol{Y}'\boldsymbol{S}]_{jl}$.

PROOF. Note that $G(\mathbf{Y},\mathbf{Y}')$ is the upper bound of Eq. (5) which clearly satisfies: (1) $G(\mathbf{Y},\mathbf{Y}')\ge\mathcal{L}(\mathbf{Y})$; (2) $G(\mathbf{Y},\mathbf{Y})=\mathcal{L}(\mathbf{Y})$.

To find the minimum of $G(\mathbf{Y},\mathbf{Y}')$, we can derive that

$$\begin{aligned}
&\tfrac{\partial G(\mathbf{Y},\mathbf{Y}')}{\partial Y_{jl}}=\\
&4\textstyle\sum_{i=1}^{n}[\mathbf{Y}'\mathbf{S}\mathbf{Y}'^\top]_{ij}[\mathbf{Y}'\mathbf{S}]_{il}\frac{Y_{jl}^3}{Y_{jl}'^3}-4\sum_{i=1}^{n}W_{ij}[\mathbf{Y}'\mathbf{S}]_{il}\frac{Y_{jl}'}{Y_{jl}}\\
&+C(2Y_{jl}-2\alpha_j\hat{Y}_{jl}\tfrac{Y_{jl}'}{Y_{jl}'})\exp(\hat{Y}_{jl}),
\end{aligned}\tag{20}$$

and the Hessian matrix of $G(\mathbf{Y},\mathbf{Y}')$

$$\frac{\partial G(\mathbf{Y},\mathbf{Y}')}{\partial Y_{jl}\partial Y_{ik}}=\delta_{ji}\delta_{lk}\Phi_{jl}\tag{21}$$

is a diagonal matrix with positive elements, where

$$\begin{aligned}
\Phi_{jl}=&12\textstyle\sum_{i=1}^{n}[\mathbf{Y}'\mathbf{S}\mathbf{Y}'^\top]_{ij}[\mathbf{Y}'\mathbf{S}]_{il}\frac{Y_{jl}^2}{Y_{jl}'^3}+4\sum_{i=1}^{n}W_{ij}[\mathbf{Y}'\mathbf{S}]_{il}\frac{Y_{jl}'}{Y_{jl}^2}\\
&+2C(1+\alpha_j\hat{Y}_{jl}\tfrac{Y_{jl}'}{Y_{jl}^2})\exp(\hat{Y}_{jl}).
\end{aligned}\tag{22}$$

Thus $G(\mathbf{Y},\mathbf{Y}')$ is convex with respect to $\mathbf{Y}$. Therefore, we can obtain the global minimum of $G(\mathbf{Y},\mathbf{Y}')$ by setting $\frac{\partial G(\mathbf{Y},\mathbf{Y}')}{\partial Y_{jl}}=0$. □

Now we turn to prove Theorem 3.

PROOF. According to Lemma 1, we can obtain the auxiliary function $G(\mathbf{Y},\mathbf{Y}')$ and thus the objective function $\mathcal{L}(\mathbf{Y},\boldsymbol{\alpha})$ in Eq. (5) has the following property:

$$\mathcal{L}(\mathbf{Y}^0,\boldsymbol{\alpha}^0)\ge\mathcal{L}(\mathbf{Y}^0,\boldsymbol{\alpha}^1)\ge\mathcal{L}(\mathbf{Y}^1,\boldsymbol{\alpha}^1)\ge\dots\tag{23}$$

So $\mathcal{L}(\mathbf{Y},\boldsymbol{\alpha})$ is monotonically decreasing. Since $\mathcal{L}(\mathbf{Y},\boldsymbol{\alpha})$ is obviously bounded below, the iterative updating process will converge to its minimum, which completes the proof. □